Path-based networking: From POTS to SDN

Malathi Veeraraghavan
University of Virginia
April 28, 2014

Talk at CUHK, Dept. of IE

Thanks to the US DOE ASCR for grant DE-SC0007341, and NSF grants CNS-1116081, OCI-1127340, ACI-1340910, and CNS-140571.

Thanks to several external collaborators (ESnet, Internet2, NCAR, UNH, UCAR, MAX, IU, Rutgers, UWisc), and several PhD/MS/UG students (http://www.ece.virginia.edu/mv/html-files/students.html)

Outline

- Problem statement
- Path-based networking technologies & Svcs
- Our research group's contributions
- Research-and-education networks (RENs)
- Potential use cases & Missing pieces
- Deployment strategy

Problem statement

- How to add a complementary global communications service to the existing IP-based Internet service widely enjoyed today?
- What is the service?
  - Some combination of rate and delay guarantees (e.g., DHL/Fedex vs. postal service)

3

Motivation

- Driven bottom-up by new technologies
  - Not top-down by applications
- To take current Internet service to the next level, however that is interpreted
- Think transportation industry

4

Outline

- Problem statement
- ➢ Path-based networking technologies & Svcs
- Our research group's contributions
- Research-and-education networks (RENs)
- Potential use cases & Missing pieces
- Deployment strategy

Types of switches:

new data-plane multiplexing/switching technolgies

Line card (mulitplexing) / Controller (admission control or not)	Circuit-switch (CS) (position-based: space, time, wavelength)	Packet-switch (PS) (header-based)
Connectionless (CL) (no admission control)		e.g., IP routers; Ethernet switches
Connection-oriented (CO) (admission control)	e.g., telephone network circuit switches, **SONET/SDH, OTN, WDM, FlexiGrid**	Virtual-circuit (VC) switches: **ATM, MPLS, Carrier Ethernet (VLAN/PB/PBB)**

- Path based (CO) vs destination based (CL) tables
  - O(number of communicating pairs) vs. O(N)

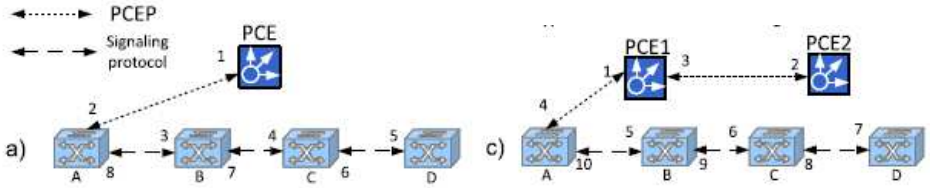
Path-based networks

need mechanism to setup/release circuits/VCs

- Control-plane solutions (IETF)
  - PCEP and RSVP-TE
- Management-plane solutions
- Advance reservation schedulers
  - Scientific computing community
  - Research-and-Education Networks (REN)
- OpenFlow and SDN

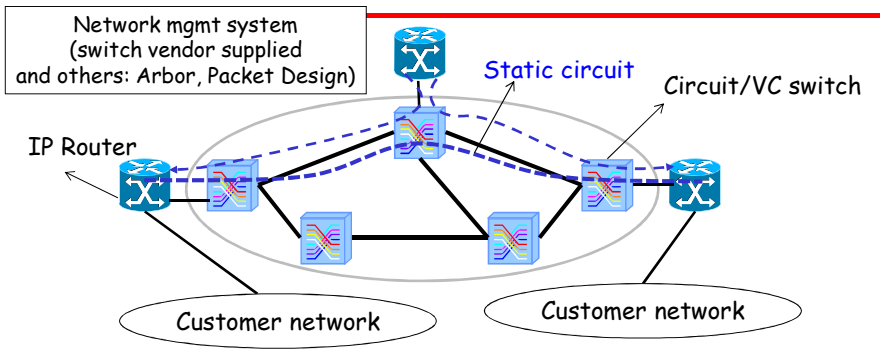
Control plane

- Path Computation Element (PCE) Communication Protocol (PCEP)
- Signaling protocol: Resource reSerVation Protocol with Traffic Engineering (RSVP-TE): setup or release the circuit/VC



F. Paolucci, et al., "A Survey on the Path Computation Element (PCE) Architecture," IEEE Comm. Surveys and Tutorials, 2013

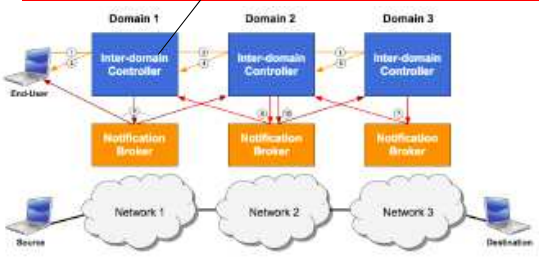
Management plane



- Fault, Configuration, Accounting, Performance, Security (FCAPS)
- Config. mgmt: create static circuits

Advance reservation service

Circuit scheduler



- Create reservation phase: endpoints, rate, duration, start time
  - Daisy-chain
  - Domain-to-Domain
- Signaling phase (provisioning phase)
  - Just before scheduled start time, IDC signals the ingress switch to initiate circuit setup
  - RSVP-TE used

- Projects:
  - DOE ESnet: On-Demand Secure Circuits and Advance Reservation System (OSCARS)
  - Internet2 ION and DYNES (Dynamic Network System)
- Protocols:
  - Inter-Domain Controller Protocol (IDCP)
  - Open Grid Forum Network Services Interface (NSIv2) standard

OpenFlow and SDN

- OpenFlow defines a standardized protocol to allow external software engines to configure forwarding tables in switches
  - Cloud computing enabled this separation
  - Cheaper switches – merchant Silicon
- Software defined network (SDN)
  - Good example: Google B4 SDN WAN: Sigcomm 2013 paper

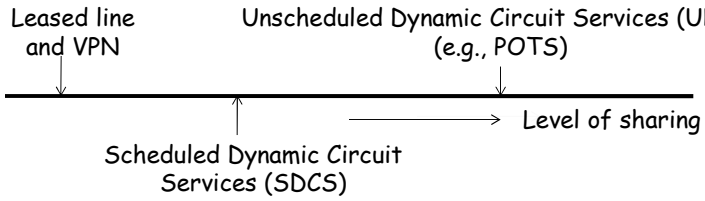
11

Outline

- Problem statement
- ➢ Path-based networking technologies & Svcs
- Our research group's contributions
- Research-and-education networks (RENs)
- Potential use cases & Missing pieces
- Deployment strategy

12

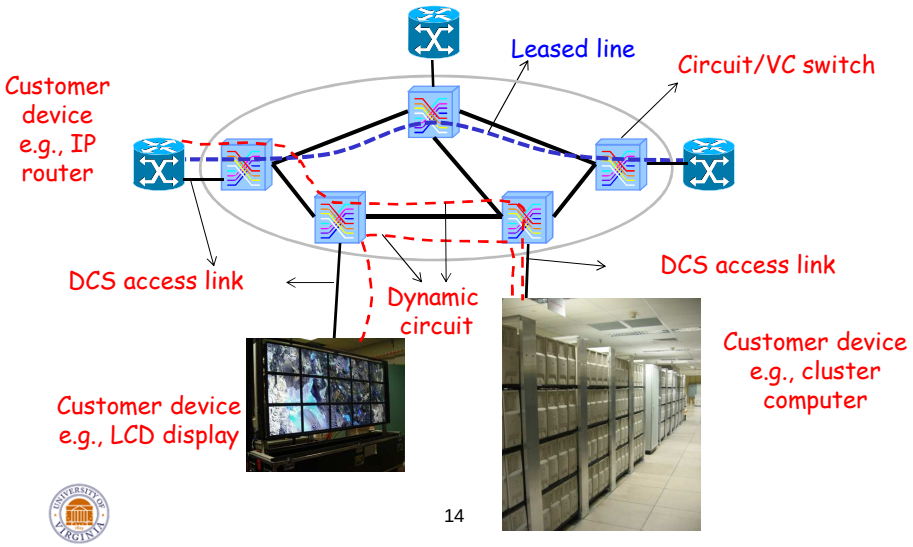
Types of path-based services



- Leased-line/Virtual Private Network
  - one contract for a static circuit: endpoints, rate, duration
- Dynamic Circuit Services:
  - Two contracts: (i) access link (ii) short-term circuits
  - SDCS: high-rate circuits (reservation system)
    - need scheduler and users/applications must specify duration
  - UDCS: ala POTS (bufferless queueing system)

13

How does DCS differ from leased line service?



14

New global service: DCS

- On global scale, today's Internet offers
  - Connectionless service
  - No delay/rate guarantees
- Leased-line/VPN service offered on global scale: but expensive, slow and inflexible (admins involved)
- Can we offer DCS on global (multi-domain) scale?
  - Needs to be heterogeneous (unlike POTS)

15

Outline

- Problem statement
- Path-based networking technologies & Svcs
- ➢ Our research group's contributions
- Research-and-education networks (RENs)
- Potential use cases & Missing pieces
- Deployment strategy

16

Our research group's contributions

- Hardware signaling implementation
- CHEETAH project
  - WAN deployment
  - Circuit TCP
- Internetworking IP and ckt/VC nets
  - gateways: servers (with disks)
  - gateways: IP routers
- Above: UDCS; some: SDCS papers

17

Hardware signaling

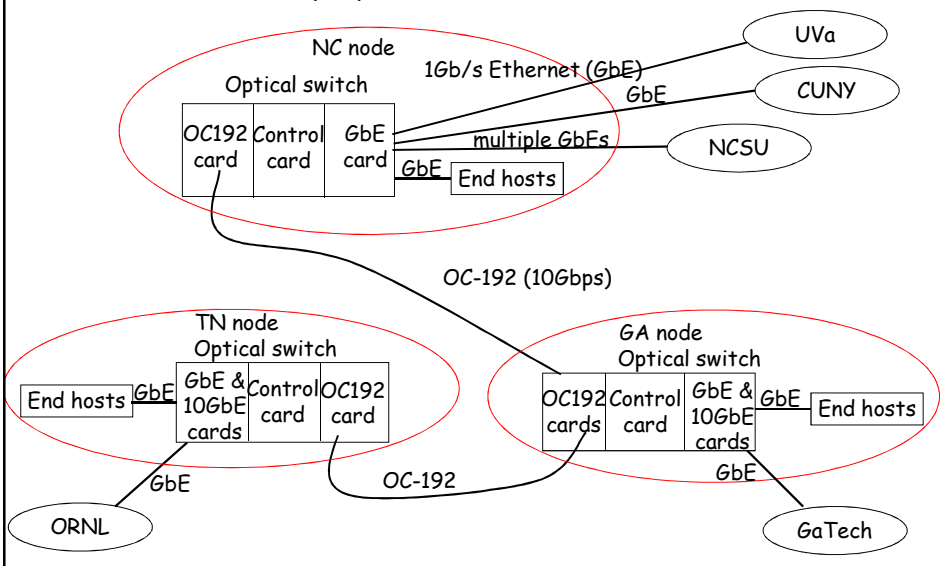
- 2001-2005 NSF Project
- Implemented RSVP-TE in hardware
  - JSAC 2005 paper (with H. Wang, R. Karri, T. Li)
- Why?
  - Intended application: file transfers; Service: UDCS
  - Files stored on disk at one host can be streamed without silences at fixed rate to remote host
  - Unlike VoIP, sounds ideal for circuits
  - But as link capacity increases, file transfer delays become small, which means setup delay can become significant part of total delay
  - With hardware signaling, setup delay reduced to one round-trip propagation delay; same as TCP connection setup delay

18

CHEETAH

- 2003-2008: Circuit switched High-speed End-to-End Transport ArcHitecture (CHEETAH): $3.5M
- Ethernet-SONET circuit-switched WAN
  - Sycamore SN16000 switches: port/VLAN mapped to SONET circuit
  - McLean, VA – Raleigh, NC –Atlanta, GA -Oak Ridge, TN
  - Circuit endpoints: computers
  - Ported UMD RSVP-TE software for Linux end hosts to communicate with Sycamore RSVP-TE built into switches
  - Unscheduled Dynamic Circuit Service
  - Application attempts circuit; if blocked fall back to IP
- Opticomm 2003 paper: won best-paper award!
- JSAC paper 2007: Experiences in implementing an experimental wide-area GMPLS network w/ X. Zhu, X. Zheng
- IEEE TPDS 2009: A hybrid network arch. for FT w/X. Fang[19]

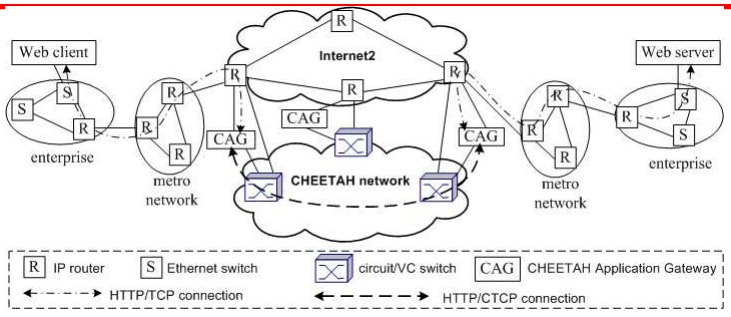
CHEETAH deployment



Circuit TCP

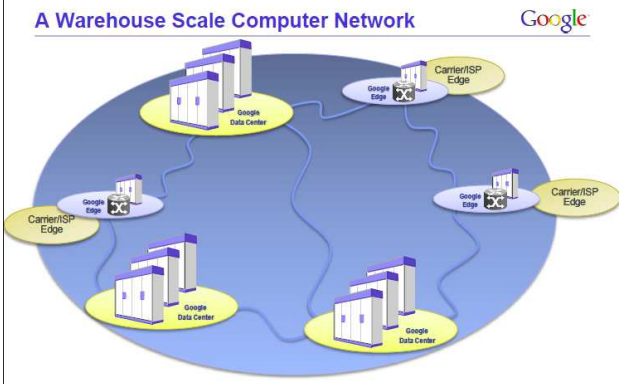
- Once a circuit/VC is set up between hosts, sender should not modify sending rate
  - CTCP eliminates congestion control
- If circuit/VC rate is full capacity of NIC, then packet-switching/TCP advantage of elastic rates is no longer useful
  - Dual NICs in hosts: one for IP, other for ckts
  - also, can use single NIC with VLANs
- IP just at end hosts for programming ease
- ICC 2006 paper w/ A. P. Mudambi, X. Zheng

Internetworking IP-ckt networks with servers



- Cloud solution! Google SDN paper concept
- ICCCN 2007: An overlay approach for enabling access to dynamically shared backbone GMPLS networks w/ X. Fang
- ICACT 2009: Internetworking circuit and connectionless networks w/ X. Fang – Circuit utilization improvement

Google B4 SDN WAN

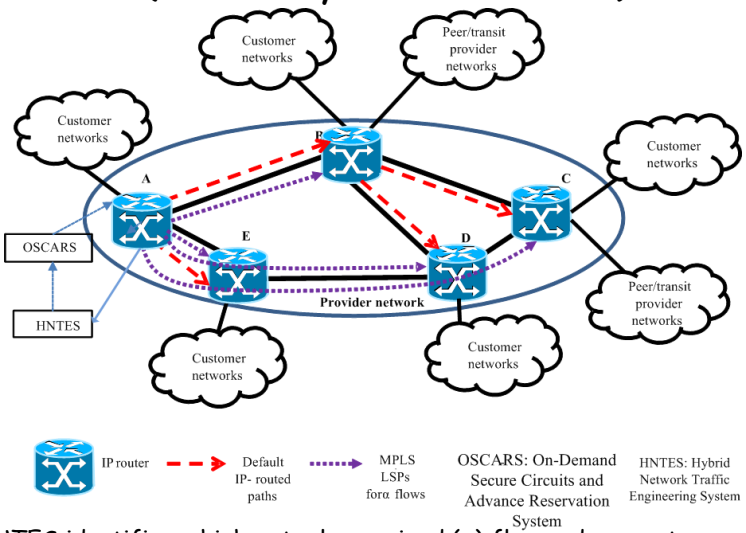


- Bikash Koley, 9/19/2013, NSF/OSA workshop

Key point: no flow enters AND leaves Google's network without stopping at a computer (in a data center)

23

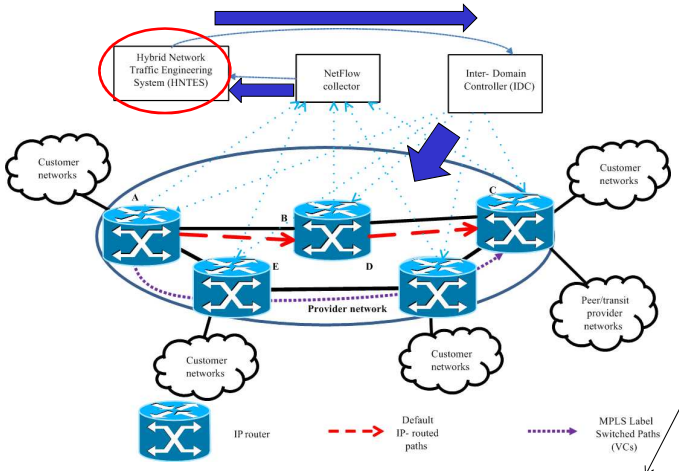
Internetworking at IP routers (to use dynamic circuits)



HNTES identifies a high-rate, large-sized (α) flow and requests OSCARS to dynamically setup an LSP and redirect α flow: challenging

24

Internetworking at IP routers (only with static circuits/VCs)



- Three steps
  - Analysis of NetFlow records from ingress routers to identify address prefixes of completed alpha flows
  - IDC creates L3 circuits between ingress-egress router pairs and configures QoS
  - IDC sets firewall filters to direct future alpha flows with matching address prefixes to L3 circuits

Aging parameter (A): age out rules corresponding to prefixes for which no alpha flows have been observed

25

Outline

- Problem statement
- Path-based networking technologies & Svcs
- Our research group's contributions
- ➢ Research-and-education networks (RENs)
- Potential use cases & Missing pieces
- Deployment strategy

26

REN deployment of SDCS

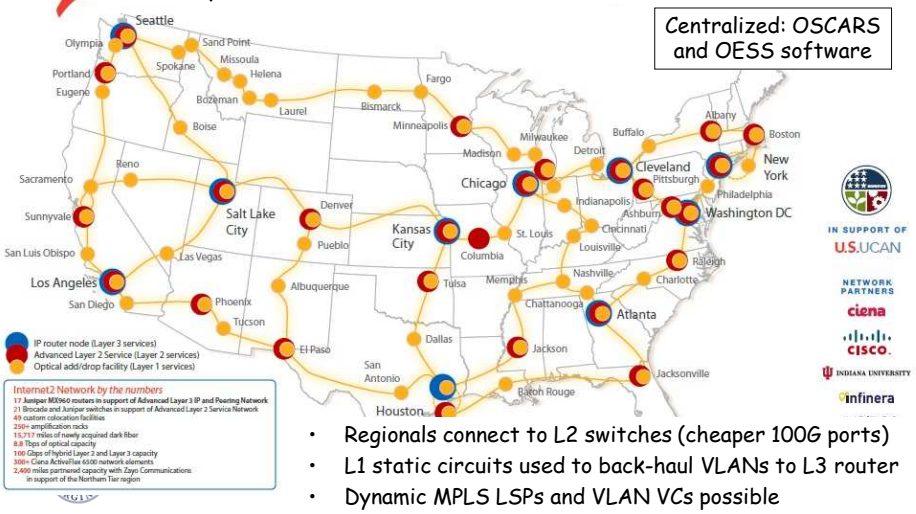
- Internet2:
  - MPLS network, VLAN network (AL2S) and L1 network
  - OESS control-plane s/w
  - DYNES: 40 university deployments
- ESnet:
  - OSCARS project: circuit scheduler
  - Large Hadron Collider (LHC) and other projects: MPLS LSPs are in use

27

Internet2 network

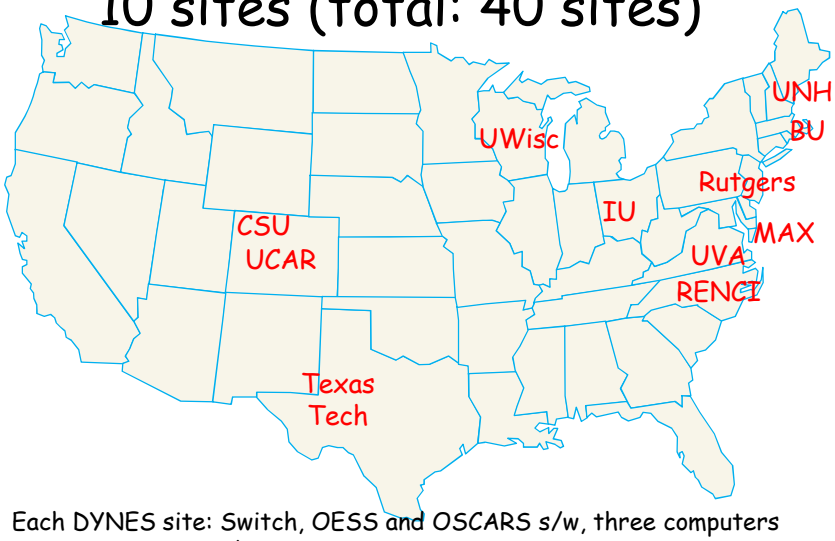
- Layer-1 and Layer-2 services: Path based
- Layer-3 service: Destination based (IP)

Centralized: OSCARS and OESS software



- Regionals connect to L2 switches (cheaper 100G ports)
- L1 static circuits used to back-haul VLANs to L3 router
- Dynamic MPLS LSPs and VLAN VCs possible

DYNES: we have logins at these 10 sites (total: 40 sites)



- Each DYNES site: Switch, OESS and OSCARS s/w, three computers
- Internet2 AL2S and ION services
- Regional services: DYNES or stitched VLANs, e.g., MARIA, FRGP

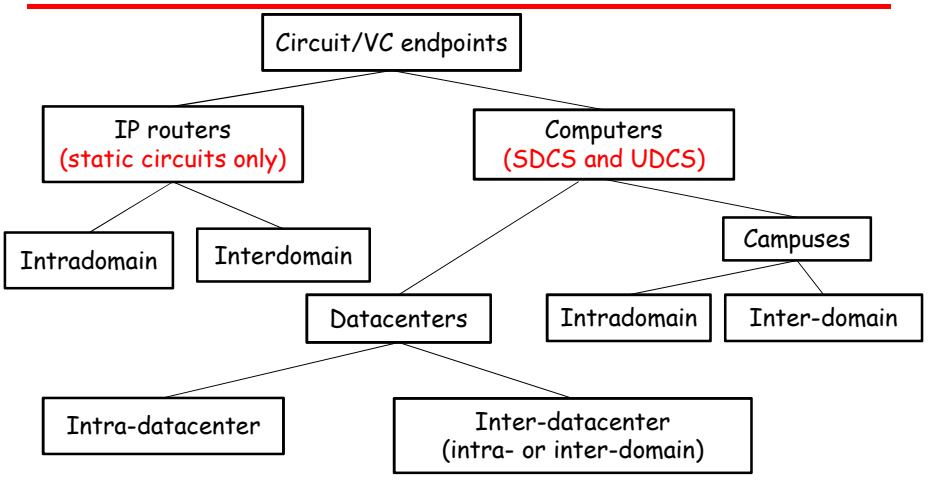
29

Outline

- Problem statement
- Path-based networking technologies & Svcs
- Our research group's contributions
- Research-and-education networks (RENs)
- ➢ Potential use cases & Missing pieces
- Deployment strategy

30

Potential use cases for circuit services



Datacenters

- Intra-datacenter
  - HELIOS, OSA, cThrough: Hybrid packet/optical circuit networks
- Inter-datacenter, intra-domain
  - Google's B4 SDN
- Inter-datacenter, inter-domain and inter-campus: more challenging
- Any intra-campus applications?

32

Outline

- Problem statement
- Path-based networking technologies & Svcs
- Our research group's contributions
- Research-and-education networks (RENs)
- Potential use cases & Missing pieces
- Deployment strategy

Missing pieces

- Inter-domain routing
  - Is BGP sufficient for SDCS/UDCS?
- Network management
  - FCAPS
  - No traceroute in L2 and L1
  - Even though OSCARS/OESS offer backup circuit provisioning, still viewed as inherently less robust than IP
- Application Programming Interface (API)
  - Sockets offer ease-of-use but need IP address configuration at ends of circuits and root access on end hosts
- Applications

Applications

- Dynamic CDN: inter-datacenter (DC)
- Scientific bulk-data movement: inter-DC
- Wide-area file systems (parallel)
- Parallel job scheduling: intra-DC
- Scientific work-flow scheduling: XSEDE
- Reliable multicast: weather data
  - static VCs: multipoint L2 VLANs

35

Deployment strategy

- Sequence
  - Intra-datacenter
  - Intra-campus
  - Inter-datacenter: REN community
  - Inter-datacenter: Commercial
  - Inter-campus (enterprise)
  - Residential

36

Types of research

		Considerations of use	
		No	Yes
Quest for fundamental understanding	Yes	Pure basic research (Bohr)	Use-inspired basic research (Pasteur)
	No		Pure applied research (Edison)

Summary

- Small but committed community
  - Most important: US federal agencies
- Would be gratifying to see widespread use of this new networking service
- Challenge lies in incremental deployment